Tackling toxic content

Mark Harwood, developer

Elastic

November 29th 2017

@elasticmark
Business drivers for tackling toxic content

Advertisers
- Withdrawing ads

Pressure groups
- Public shaming

Government
- Fines, legal restrictions

Consumers
- Reputational damage, loss of audience

Pressure groups

Hate speech

Fakes news

Extremist videos

Toxic content
How?
Two approaches:

- **Proactive**
  - Root out content before it gathers an audience

- **Reactive**
  - Respond to complaints from the audience

How do your staff determine what is “toxic”?

Whose opinions do you trust?
Proactive challenge

How do we determine what is toxic?
Content based analysis is hard

- **Parsing is hard** - content is often binary e.g. audio or video
- **Limited metadata** - lack of descriptions or keywords
Easier to examine activity around content

- Reuse the basis of recommendation engines - people who liked X also like Y
Recommendations recap: MovieLens data

```
{
  "movie": [260, 500, 1080...]
  "user": 8353
}
```

http://files.grouplens.org/datasets/movielens/ml-10m-README.html
Random samples should hold no surprises

• 17% of all people like “Forrest Gump”
• In a random sample of people, 17% of them will also like “Forrest Gump”

Dull. But in non-random samples something interesting happens.....
Non-random sample: people who liked “Talladega nights”

Find all people who liked movie #46970

In the set of “Talladega-likers”, 20% of them like “Anchorman”

..a huge uplift in popularity from the norm!

Summarise how their movie tastes differ from everyone else
Proactive demo
Reactive challenge

Whose opinions do we trust?
Allow end users to report toxic content
BUT - some user reports, like some content, can be questionable
Review fraud is a thing

- **Positive reviews** - “shill” or “sock puppet” accounts are used to artificially inflate the reputation of sellers in a marketplace

- **Negative reviews** - fake accounts or mob-rallying is used to sabotage the reputation of an innocent party.

Tell-tale signs of collusion might include:

- A common IP address or user agent
- A common "hit list" of items being flagged
- A common phrase used in feedback
- The same time-of-day when logging requests
- The same site join-date
Components of a fraud detection stack

Ingest
- Cleansing, enriching, normalisation

Linking
- Entity resolution, filtering

Risk-scoring
- Graph exploration, anomaly detection, scoring

Investigation
- Task lists, case management, visualisation

Outcomes
Bad actors make strange shapes

It is hard for identity manipulators to avoid reusing resources (IP addresses, join dates, subject lists, phrases, time). Fraudsters generate too many “coincidences”. Use the Graph API to gather related data then raise alerts on anomalies.

Responding to alerts

Ingest → Linking → Risk-scoring → Investigation

Kibana with the Graph plugin allows investigators to examine details behind alerts.

Demo